The Biomedical Information Explosion:

From the *Index-Catalogue* to MEDLARS

BY STANLEY JABLONSKI, *Head*

Index Section, Bibliographic Services Division
National Library of Medicine
Bethesda, Maryland

ABSTRACT

This study demonstrates the growth of medical periodical literature by means of the increase of titles indexed in the *Index-Catalogue* and in *Index Medicus,* with reference to other biomedical-related titles not covered in these publications. Stress is laid on the fact that information outside the biomedical area is needed by the researcher. Problems of information retrieval, with the specific needs of the user in mind, are discussed, together with the particular problems of indexing which this specificity raises.

THE term "information explosion" is often used rather inaccurately to dramatize the increase in the production of scientific literature. It suggests that quite recently a sudden and unexpected outburst occurred in publishing activity, whereas the available evidence indicates that the literature has been growing at a rapid but essentially unchanging pace for more than a century, doubling in size every twenty-five years. Licklider sees the information explosion as follows:

> What is happening in scientific and technical communication is more closely analogous to a flood than to an explosion. The water has been rising slowly—at a progressively increasing rate, but slowly, for a long time, and now it is overflowing the banks in several areas. Although the rise has been gradual, the overflow is sudden and dramatic (1).

No accurate figures on the volume of biomedical literature published over the years are available, partly because few institutions held sufficiently complete collections, but chiefly because there has been little agreement on the scope of the biomedical sciences and on what is and is not of importance to the profession.

Those records which are available, however, indicate that ours is not the first generation to have been overwhelmed by the volume of published material. Rogers and Adams report that as early as 1878 indexing back files and current journals had accumulated 400,000 unpublished subject entries (2). Coverage of the *Index-Catalogue* also shows that the rapid growth of scientific literature can hardly be considered a new phenomenon:

1880— 2,300 titles (3)
1916— 4,700 titles (4)
1948—10,499 titles (5, 6)

Our inability to cope effectively with the steady growth of the literature is well illustrated by the constantly increasing backlog of unpublished references to periodic literature in the *Index-Catalogue* (2):

	Published Citations	*Unpublished Citations at End of Period*
1880–1895	679,669	212,000
1895–1920	848,285	1,000,000
1920–1950	1,337,247	1,750,000

Using as a base the 2,300 (3) titles of journals indexed in the *Index-Catalogue,* which normally covered about half the available literature, one can estimate that in 1880 there were between 4,500 and 5,000 titles within the broad scope of the biomedical sciences.

In 1969 somewhat fewer than 20,000 serial titles are held by the National Library of Medicine. How many of these titles are of substantial interest to the medical and biomedical professions remains an unanswered question. If one were to apply the criteria of the editors of the *Index-Catalogue,* thereby judging important all journals containing substantive articles in any specialty of medicine, as well as those in various fields of science, technology and philosophy with direct or potential application to human health and disease, half or more of all existing

publications would be in scope. If, on the other hand, the more stringent criteria of the *Current List of Medical Literature* (which replaced the *Index-Catalogue* in 1950) were used, only those titles "selected on the basis of proportion of original work reported in the journals, and on the pragmatic basis of journals most commonly held by American libraries" would have been included, reducing the scope to 1,225 titles in 1950 (2) and to 1,700 in 1959 (7). Excluded would be journals of lesser quality, those not commonly available in American libraries, and publications from various paramedical and peripheral disciplines already covered by nonmedical special bibliographies such as *Chemical Abstracts*, *Biological Abstracts*, or *Psychological Abstracts*.

When the *Current List* was in turn replaced by the *Index Medicus* in 1960, the scope was once again redefined and expanded to include a number of paramedical fields and (like the *Current List* but unlike the *Index-Catalogue*) the criterion of quality was retained. By 1969 the *Index Medicus* included about 2,400 titles (8).

When comparing the nature of the 2,400 titles in the *Index Medicus* (8) with the approximately 3,300 titles held by the NIH Library (9) which, one must assume, reflects the needs of medical and biomedical researchers, a similarity in coverage is apparent in most specialties, with the exception of the basic sciences which are more broadly represented in the NIH Library. Whether or not coverage of the biomedical periodical literature in the *Index Medicus* and the NIH Library includes all important publications may be open to question, but the fact remains that their coverage does support a conviction of many librarians: namely, that most users rely chiefly on a relatively small number of journals.

The scope of biomedical sciences may be defined as a sum of disciplines which, in addition to the traditionally recognized medical and biological specialties, includes those aspects of science, technology, and all other fields which contribute to the knowledge of human health and disease. Its boundaries are never static, being constantly influenced by new discoveries, research trends, social needs, availability of resources, changes in orientation within disciplines, and a host of other factors. Each field relates to other fields, both within and outside the biomedical sciences, but the relationship assumes different forms in each instance, each request, each bibliography, and each study, and reflects the vantage point of the specialty, the purpose of this study, the era, the state of knowledge, or the requirements of individual researchers, including their prejudices. Though man is the central theme in the pattern of the biomedical sciences as a whole, with all else subordinate, he may be relegated to the periphery or even eliminated in some areas, especially in those which, while having no apparent relevance to human health, are nevertheless a part of larger areas of problem solving.

There is no rule, moreover, which states that any subject field, especially one in the periphery, must be assigned a permanent place in the biomedical sciences. To use a somewhat extreme example, metallurgy is not considered a biomedical specialty; yet, the discovery that a particular processing technic may increase the toxicity of certain metals is of definite importance to the medical profession. Nevertheless, this does not mean that all literature on the technic and the toxic metals thus fabricated should automatically be covered by biomedical information services, nor that coverage should go on after the offending process has been discontinued.

In his *Information Retrieval Systems* Lancaster outlines the basic functions of information retrieval systems as follows:

> An information retrieval system does not inform (i.e., change the knowledge of) the user on the subject of his inquiry. It merely informs him on the existence (or nonexistence) and whereabouts of documents relating to his request (10).

One implication of this traditionally accepted concept is that judgment of the pertinence, whether because of the subject or quality, of a document is the sole prerogative of the user, and the role of the information specialist is limited to assistance, without analysis, editorializing, or commenting on quality or value, in locating documents which are likely to contain material of potential interest to the user. Even the editors of the *Index-Catalogue*, themselves distinguished physicians and scholars, were especially careful not to infringe on this prerogative of the user.

One consequence of this traditional approach was the establishment of scope policies of most medical bibliographies, most notable of which

was the *Index-Catalogue*, which seldom allowed for criteria of quality in selecting journals for indexing, and which included many subject areas that were quite peripheral to medicine and biology. Another was the use of relatively broad subject headings; this offered the user large selections of documents on a variety of subjects and specialties among which to browse, allowing him to choose at his convenience those references which he himself considered important and to reject the rest. With this approach, more than half of all material printed would have been considered important to the biomedical profession.

The basic function of an information service is still to inform the user about the existence and whereabouts of documents relating to his subject of study. This may be as simple as pointing to a shelf where a particular textbook is located or as complex as providing a listing of the literature on pathological changes of the corpora flava in inborn errors of metabolism transmitted as a sex-linked trait and involving amylo-1,6-glucosidase deficiency. The current trend, however, is toward greater specificity. A present-day requester, whether because of his greater specialization, sophistication, lack of time, or a combination of several factors, is not interested in browsing through literature for which he has no use. Instead, he is likely to ask specific questions and expect specific information on literature directly pertaining to his field of study. It is no longer enough to tell him, "Here are documents which are likely to contain material of possible interest to you. Now decide what you want and discard the rest." You must tell him, "Here is the exact information you asked for."

In a conventional bibliography the degree of specificity with which documents are indexed is nearly always proportional to the number of entries.

For example, if we were to accept ten citations per heading or a heading-subheading combination as a desirable number, a bibliography containing 1,000 entries would be expected to have 100 headings or any combination of headings and subheadings which would result in the product of 100. In a bibliography covering the same subject but containing 10,000 entries, the number of headings and subheadings and, consequently, the degree of specificity would be ten times greater.

Accordingly, in the *Index-Catalogue* the degree of specificity of indexing was fixed in relation to the number of articles indexed. Through a liberal use of subheadings and cross references which is unmatched today, the editors of the *Index-Catalogue* were able to achieve a much greater specificity than its 2,000 or so headings and restrictions of the system would seem to justify, but even their skill could not override the deficiency of a system not geared to the degree of specificity now demanded. In the *Index-Catalogue*, much as in other bibliographies of the era, the degree of specificity was determined by the volume indexed rather than the needs and demands of the user.

Furthermore, bibliographies such as the *Index-Catalogue* are unsuited to high-specificity information retrieval because subject headings, usually one- or two-word terms, cannot in themselves express a total idea. The word *cell*, for example, may be a totally different meaning to an engineer, or a political scientist, or to a cytologist; in a biomedical bibliography its meaning is reasonably well defined; it expresses only a broad concept. Even though the word *cell* represents a finite concept, its specificity as a subject heading is relatively not much greater than that of terms representing whole organs or systems, such as *liver*, *brain*, *stomach*, or *gastrointestinal system*, just as the word *toothpick* is not more specific than *tree*, nor *snowflake* than *snow* or *winter*.

Contrary to widely held opinion, the specificity of individual subject headings seldom in itself insures the specificity of information retrieved. Moreover, in the quest for specificity it is not only the specificity of the subject fields that we are after but also specificity of the various relationships which these fields may assume. It is quite unlikely that a requester would ever ask for all the literature on *cell nucleus* (which, incidentally, accumulated 5,190 references on MEDLARS tapes from 1964 to 1969), any more than he would seek all the literature on *stomach* (3,875 references), *liver* (25,000 references), or *brain* (11,200 references). He is more likely to request the literature on the effects of gamma rays on hydroxylase activity in liver cell nuclei in the adrenalectomized male rat.

We are now totally committed to the concept of specificity in information services, but we

still cannot agree on what degree of specificity we should aim for; the old question of what is important to the user and what is not remains to be answered. The main problem here arises from the fact that information services always work prospectively, providing for the future while using past experiences as their guide. In an era when some types of scientific information are considered outdated after only two years, the business of predicting future needs of the scientist can be quite tricky.

The literature on cancer, as an example, may well be within the spheres of interests of various specialists, though we expect an epidemiologist to view it from a vantage point unlike that of a biochemist or a surgeon. Let us say that we know some types of gastric cancer are more prevalent among certain ethnic groups; we may prepare for queries in this area by providing information on the race, socioeconomic status, and geographic distribution of patients, together with pertinent cancer data. However, let us suppose that future studies link breast cancer to the wearing or not wearing (in some modern circles) of certain undergarments, or that they show a relationship between the consumption of martinis and cancer of the bladder, or point to eye color as a significant factor in the occurrence of some cancers. Not having previously considered the lingerie, drinking patterns, or eye color of cancer patients as important, or even relevant data, we will thus be completely unprepared to provide information on these new and conceivably important aspects of the epidemiology of cancer.

The concept of the word "information" within the context of information services has been changing slowly but definitely for the past several decades. Whereas before we were merely informing the user about the general subject of a document, we are now concentrating on all available data within the document. We are seldom interested in a single subject, moreover, but usually in entities and ideas expressed through the various specific relationships of individual topics discussed in the document.

The demand for specificity and selectivity recently imposed on information services is well reflected in Elwin's report on the activities in 1969 of the MEDLARS Center at Karolinska Institutet in Sweden. Out of about 400 literature searches performed at the institute, fewer than ten citations per search were considered pertinent by 39 percent of requesters and fewer than twenty by an additional 20 percent (11).

Since there were more than 1,000,000 citations on MEDLARS tapes at the time, more than half of the Karolinska requesters appear to have been interested in less than 1/50,000 of the available literature and more than one-third in less than 1/100,000. The question, therefore, that Elwin's report seems to be begging is: "Why is there so little in the literature on specific subjects at a time when we are supposed to be in the midst of an information explosion?"

The crux of the matter appears to be that, even though there are claims suggesting as many as 20,000 active journals which have been estimated to contain as many as 2,000,000 new articles that may be considered within the biomedical sciences broadly outlined, they represent a wide variety of separate subjects and serve an equally large variety of separate specialties and disciplines whose interest might often differ to the point of total divergence. A practical nurse and a biophysicist, both biomedical specialists, have little in common; their interests in the literature are wholly disparate; yet they draw information from the same pool and use many of the same information services. Even those specialists whose background and fields of study are quite similar nearly always approach the literature differently.

The chief reason for this phenomenon lies in the development of information services. The earlier user was obliged to scan hundreds of irrelevant articles to locate a few relevant ones; this falsely implies that his spheres of interest were much wider than was actually so. Now, with improved information technology and, particularly, with the user's greater specialization and selectivity, in conjunction with the increase in the volume of the literature, only one out of 100,000 recent articles (and one out of several million of the total number of available articles) is likely to be considered pertinent by the user.

One often hears about information being measured in terms of feet of computer tape, trillions of alphanumeric characters, or hours needed to read all published scientific works. This perhaps dramatizes the problem well enough to convince anybody that an information explosion has really occurred. But why should we think of knowledge and information as having form and substance measurable by

conventional yardsticks? Why, above all, should anybody ever be interested in reading everything scientific, any more than a student of current events should be interested in reading cover-to-cover every newspaper in the world?

To summarize, therefore, the author believes that the information explosion does not represent a sudden increase in the growth of the literature, nor does he believe it represents a sudden increase in the knowledge contained in the literature. Rather, the suddenness involved relates to the availability of information previously inaccessible—information now available because of advances in information services brought about by a complex of factors. Included among the latter are the introduction of automatic data processing, more specialized needs of users (along with a resultant demand for greater specificity and selectivity), and the expanding scope and growing volume of the published literature of the biomedical sciences.

REFERENCES

1. LICKLIDER, J. C. R. A crux in scientific and technical communication. Amer. Psychol. 21: 1044–51, 1966.
2. ROGERS, F. B., AND ADAMS, S. The Army Medical Library's publication program. Texas Rep. Biol. Med. 8: 271–300, 1950.
3. Alphabetical List of Abbreviations of Titles of Medical Periodicals Employed in the *Index-Catalogue*. Index-Catalogue of the Library of the Surgeon General's Office, United States Army. Washington, Government Printing Office, 1880, Ser. I, Vol. 1, p. 1–126.
4. First Addition to the Alphabetical List of Abbreviations of Titles of Medical Periodicals in the Twenty-first Volume, Second Series of the Index-Catalogue of the Library of the Surgeon General's Office, United States Army, Washington, Government Printing Office, 1918, 231 p.
5. List of Abbreviations for Serial Publications Referred to in the Fourth Series of the Index-Catalogue. Washington, Government Printing Office, 1948, 138 p.
6. Supplementary List of Abbreviations for Serial Publications Referred to in the Fourth Series of the Index-Catalogue. Washington, Government Printing Office, 1955, 79 p.
7. List of Journals Indexed. Current List of Medical Literature. Washington, National Library of Medicine, 1959, 20 p.
8. List of Journals Indexed in Index Medicus. Bethesda, National Library of Medicine, 1969. 99 p.
9. Periodicals Currently Received in the NIH Library. Washington, U. S. Department of Health, Education, and Welfare. Public Health Service. National Institutes of Health, 1969. 232 p.
10. LANCASTER, F. W. Information Retrieval Systems: Characteristics, Testing and Evaluation. New York, Wiley [c1968], 222 p.
11. ELWIN, C. E. MEDLARS, Karolinska Institutet, Monthly Recurring Demand Search Service. Stockholm, 1969. (Unpublished report)